78509_Auto_Edited.docx

WORD COUNT	5317	TIME SUBMITTED	15-SEP-2022 10:52AM
		PAPER ID	90525574

Name of Journal: *World Journal of Gastroenterology*

Manuscript NO: 78509

Manuscript Type: ORIGINAL ARTICLE

Basic Study

Hybrid XGBoost model with Hyperparameter Tuning for prediction of Liver disease with better accuracy

Hybrid XGBoost model with Hyperparameter Tuning for prediction of Liver disease with better accuracy

Surjeet Dalal, Edeh Michael Onyema, Amit Malik

Abstract

BACKGROUND

Liver disease is a leading cause of mortality in the United States and is regarded as a life-threatening condition not only in this nation but all across the world. It is possible for people to develop liver disease at a young age.

AIM

More than 2.4% of all fatalities in India occur as a result of liver disease each year. In addition, the early indications of liver illness make it harder to diagnose. When it's too late, the warning signals are always there. As a result, a more accurate and dependable automated method is needed to detect liver illness at an early stage.

METHODS

The illness can be predicted using certain machine learning algorithms. Predicting liver illness with more precision, accuracy, and reliability can be accomplished through the use of a modified XGBoost model with hyperparameter tuning in comparison to CHAID & CART models.

RESULTS

The CHAID & CART model have achieved the accuracy level 71.36% & 73.24% respectively. With the help of the proposed model, the accuracy level has been achieved up to 93.65%.

CONCLUSION

As a result, it helps patients to predict liver-related disease acutely by identifying the disease's causes and suggesting better treatment options.

Key Words: Liver Infection; Machine Learning; CHAID; CART; Decision tree; XGBoost; Hyperparameter Tuning

Dalal S, Onyema EM, Malik A. Hybrid XGBoost model with Hyperparameter Tuning for prediction of Liver disease with better accuracy. *World J Gastroenterol* 2022; In press

Core Tip:

This manuscript proposed the hybrid XGBoost model for prediction of liver disease. This model has been designed by optimizing the hyperparameter tuning with help of Bayesian optimization. But these models are not found accurate in predicting liver disease among the Indian patients. It consists of different physical health status i.e. Level of Bilirubin, Direct Bilirubin, Alkaline Phosphatase, Alamine Aminotransferase, Aspartate Aminotransferase compound, total Proteins, albumin as well as Globulin. This work is aimed to design a more accurate machine learning model in liver disease prediction for maintaining good health level of citizens.

INTRODUCTION

An organ having the size of a football, the liver processes blood from the digestive tract. It is located on the right side of the abdomen, directly below the rib cage. The liver plays a critical role in the digestion of food and the elimination of toxins from the body. It is possible that liver illness is passed down via families (genetic). Viruses, alcohol, and obesity are all known to harm the liver, which can result in liver disease. Conditions that affect the liver can lead to scarring (cirrhosis), which can eventually lead to liver failure, a life-threatening condition [1]. The liver may be able to recover if therapy is started early enough. The patient will also be recommended one or more tests in order to correctly identify and determine the cause of liver illness. These may include:

- **Blood tests**: Human blood is tested for the presence of liver enzymes by using liver enzymes. Additionally, a blood-clotting test known as the International Normalized Ratio (INR) is used to measure liver function (INR). Problems with liver function may be the cause of elevated levels[2].

- **Imaging tests**: Ultrasound, MRI, and CT scans can all be used to examine a patient's liver for damage, scarring, or malignancies. Ultrasound fibroscan can be used to measure scarring and fat accumulation in the liver, among other things.
- **Liver biopsy**: Small amounts of tissue are removed from the liver with the use of an ultra-thin needle during the biopsy procedure. The tissue is examined for any indications of liver illness.

The heart, lungs, belly, skin, brain, cognitive function, and other elements of the nervous system can all be affected by liver disease. For a physical exam, the complete body may be examined. Blood tests can be used to determine the extent of liver inflammation and how well the organ is functioning [3]. There are causes of liver disease. Cirrhosis can be caused by alcohol misuse. Nonalcoholic fatty liver disease and chronic hepatitis B and C are other possible causes. Other issues may be drug overdoses. Acetaminophen and other drugs might damage the liver if the patient use them in large doses. Keep in mind that acetaminophen may appear in several medications the patient take, so pay attention to the dosage directions on the label. Having too much fat in the liver is known as nonalcoholic fatty liver disease (NAFLD). The liver may become inflamed as a result of the excess fat. Nonalcoholic steatohepatitis is a kind of NAFLD that affects the liver (NASH). The liver is inflamed and damaged as well as full of fat. There is a risk of cirrhosis and liver scarring as a result of this medication[4]. Dire complications of liver disease include acute liver failure, when the patient does not have a long-term liver illness, but the liver shuts down in a matter of days or weeks. Overdosing on acetaminophen, becoming infected, or using prescription medicines can be the cause for this. Another type is Hepatic cirrhosis which is the accumulation of scar tissue. The more the healthy liver tissue is lost to scarring, the more difficult it is for the liver to perform its functions. It may not operate as well as it should in the long run.

The so-called "liver blood tests" are generally a good indicator of liver damage since they show anomalies in specific blood tests (for example, ALT, AST, and alkaline phosphatase enzymes). It is common to refer to all of the liver blood tests as "liver function tests." This does not mean that liver dysfunction is the cause of all abnormalities in the blood tests, such as, high bilirubin, lower than normal levels of albumin, and a prolonged prothrombin time. In addition, abnormalities in other liver blood tests may be indicative of liver damage. Hepatitis viruses, for example, may boost the levels of the ALT and AST enzymes in the blood, causing them to leak into the circulation.

This paper aims to fulfil various objectives. First objective of this paper is to identify the symptoms of the liver disease and its impact on the patient's body. Then, it presents the study of various machine learning approaches for predicting liver disease and evaluates the performance of decision-tree algorithms in prediction of liver disease. Next, the paper proposes a modified XGBoost model with hyperparameter tuning mechanism and, finally, it validates the performance of the proposed model by comparing it with more traditional decision tree based models.

RELATED WORK

A lot of work has been performed in the field of Liver disease prediction. Liu et al.[1] used shallow shotgun metagenomic sequencing of an enormous populace based partner (N > 7,000) with ~15 long stretches of follow-up in blend with AI to examine the prescient limit of stomach microbial indicators separately and related to traditional gamble factors for episode liver sickness. Independently, traditional and microbial variables showed practically identical prescient limit. In any case, microbiome expansion of traditional gamble factors utilizing AI essentially worked on the presentation. Also, infection free endurance investigation showed essentially further developed delineation utilizing microbiome-expanded models.

Kang et al.[2] upheld the expected clinical legitimacy of stomach metagenomic sequencing to supplement ordinary gamble factors for expectation of liver infections. Examination of prescient microbial marks uncovered beforehand obscure taxa for liver illness, as well as those recently connected with hepatic capability and infection. This study

Survarachakan et al.[3] intended to develop a few brain network models utilized for breaking down the physical designs and sores in the liver from different imaging modalities, for example, processed tomography, attractive reverberation imaging and ultrasound. Picture examination undertakings like division, object discovery and arrangement for the liver, liver vessels and liver sores were discussed.

Lysdahlgaard et al.[4] researched and 91 papers were sifted through for the study in view of the subjective pursuit including diary distributions and gathering procedures. The papers audited in this work were gathered into eight classifications in view of the approaches utilized. The presentation was estimated in light of the dice score for the division, and exactness for the arrangement and recognition assignments, which were the most regularly utilized measurements.

Liu et al.[5] expressed that the determination and treatment of liver illnesses from registered tomography (CT) pictures was an imperative work for division of Liver and its growths. Because of the lopsided presence, fluffy lines, different densities, shapes and sizes of sores division of liver and its growth, it is a troublesome work. The authors actually focused on profound gaining calculations for portioning liver and its cancer from stomach CT, there on limiting the time and energy utilized for a liver sickness finding. The calculation utilized here depended on the altered ResUNet design.

Yang et al.[6] utilized AI calculations to develop a gamble expectation model for liver cirrhosis disorder with hepatic encephalopathy, showing that the conservative techniques of machine learning used previously failed to identify some classes for this disease because of unbalanced data which is obtained generally in such cases. The work showcased that weighted random forest work with better accuracy as compared to weighted SVM or logistic regression method. The authors gathered clinical information from 1,256 patients with cirrhosis and performed preprocessing to extricate 81 elements from these unpredictable information.

Haas et al.[7] fostered a stomach MRI-based AI calculation to precisely gauge liver fat (relationship coefficients, 0.97-0.99) from a reality dataset of 4,511 moderately aged UK Biobank members, empowering evaluation in 32,192 extra people. 17% of members had anticipated liver fat levels demonstrative of steatosis, and liver fat could never have been dependably assessed in view of clinical factors like BMI. A vast affiliation investigation of normal hereditary variations and liver fat reproduced three known affiliations and distinguished five recently related variations in or close to the MTARC1, ADH1B, TRIB1, GPAM, and MAST3 qualities (p < 3 3 108). A polygenic score incorporating these eight hereditary variations was emphatically connected with future gamble of ongoing liver sickness. The authors show that their imaging-based AI model precisely gauges liver fat and might be valuable in epidemiological and hereditary investigations of hepatic steatosis.

Gómez-Gavara et al.[8] introduced the examination and correlation of information of patients with liver brokenness by gathering data on liver illness and gathering information for choice in information mining. The Liver Disorders Data Set (UCI Machine Learning Repository) was utilized to contrast the 359 patients and liver infection. On account of the correlation results, Tree Random Woods offers the most dependable benefit.

Shen et al.[9] featured that Liver infections are a steadily developing worldwide issue. Liver fibrosis or liver steatosis were many times noticed going with liver infections. Presently, transient elastography is in many cases utilized as a painless device to evaluate liver wellbeing however the relating gear was similarly mind boggling and costly.

Man et al.[10] presented the technique Qualitative Gene articulation Activity Relationship (QGexAR) which is a SVM based method to predict the chances of a chemical becoming cause for liver injury if it is used on a person. The drugs may use the chemicals which can be harmful and injurious to liver. Therefore, tools are required to find out the effect of a drug on liver. Currently available tools are a challenge, so, the attempt was to develop a model which can overcome the problems. Their proposed model worked with an accuracy of 72% during training and produced results with 95% accuracy during validation.

Talari et al.[11] expressed that the quantitative MRI metric, T1, can be utilized to describe fibroinflammation in the liver; in any case, the T1 esteem alone could not separate among fibrosis and aggravation. The authors assessed the expected utility of old style

AI procedures (K-Nearest Neighbors, Support Vector Machine and Random Forest) to resolve this issue involving data in the T1 map. The authors additionally thought about using various strategies to mitigate the impacts of class awkwardness. Irregular Forest with Adaptive Synthetic Sampling was better than mean T1 in ordering fibroinflammation.

Forlano et al.[12] worked on quantifying the problems associated with fatty liver in population who are non-alcoholic (NAFLD) with the help of machine learning model on 246 patients' record. These patients were suffering from NAFLD, as proved by biopsy reports. The study used interclass correlation coefficient (ICC) as performance measuring parameter and showed that the parameter obtained value of more than 0.90 for each quantified problem like inflammation, steatosis, ballooning, etc, which means that their method was fairly efficient in correlating the patient to appropriate class of problem.

Li et al.[13] expressed that Liver illness is one of the vital reasons for large number of deaths in the nation and is viewed as a perilous sickness, anyplace, yet around the world. Liver sickness can likewise affect individuals from the get-go in their life. More than 2.4% Indians passes away yearly due to liver problems. It was likewise challenging to identify liver illness because of gentle side effects in the beginning phases. The study attempts to predict the liver disease via multiple machine learning techniques in which SVM classifier shows best results.

Li et al.[14] uses binary classifiers and fine tune their performance to identify the cancer of a particular part area in liver by recording the endogenous fluorescence through a fibre optical device. The main of this study was to identify the best performing classifier through a series of experiments.

Byrne et al.[15] presented the correlation and examination of information of patients with liver brokenness by gathering data on liver illness. The study showcases that how the random forest tree mechanism can be used to predict the population with liver dysfunction with better accuracy as compared to the techniques like OneRrule, tree decision stump, ReptTree.

Nitski et al.[16] aimed to expose the risk of patients who had liver transplant in recent past and experimented with four different machine learning algorithms. The implemented techniques are compared with the performance of logistic regression model using the AUROC, area under curve method, as metric. Two separate datasets were used from sources, Scientific Registry of Transplant Recipient (SRTR) and University Health Network (UHN) with 42146 and 3269 observations respectively. The study showcased that the Transformer deep learning model outperformed the logistic model for both the datasets and suggested that the model can be very helpful in making critical intervention between life and death, of patients with transplant history, due to reasons such as cardiovascular diseases, cancer, etc[17].

From the work presented in these papers, it can be deduced that there is a scope of improving the accuracy level of traditional machine learning algorithms so this fact motivates us to propose the modified XGBoost algorithms with hyperparameter tuning.

MATERIALS AND METHODS

Dataset

There is a constant rise in the population encountering liver related diseases due to unhealthy environment for breathing, excessive alcohol intakes, contaminated elements in the diet, improper use of over the counter drugs, etc[18] (Table 1).

Table 1 depicts description of the Dataset used for the experimental purpose. The dataset used in this study constitutes 416 people with liver problems, and 167 with no such history, collected from the state of Andhra Pradesh, India[19]. The dataset divides the population into two sets, depending upon whether the patient is suffering from disease or not, and this binary information is recorded in the attribute “is_patient”. The effort is to correctly predict the value of this field so that the task can be automated and eased for medical personnel. The dataset contains record for both male and females[20].

Data pre-processing

Data preparation is the most critical step before running various machine learning models. Machine learning does not perform as expected when datasets are not handled properly. It's possible that the performance of a machine learning model during training and testing will diverge. Data errors, noise, and omissions can all contribute to this. Prior to comparing data, pre-processing removes any duplicates, anomalies, and other inconsistencies. This ensures that the findings are more accurate[21].

Handling the Null/Missing Values

Mean/median or mode are used to fill in the blanks in the dataset, depending on the type of data that is missing[22]:

- **Numerical Data**: Whenever a numeric number is omitted, a mean or median value should be used instead. The outliers and skewness contained in the data pull the average and mean values in their respective directions, hence it is preferable to impute using the median value.
- **Categorical Data**: When categorical data is lacking, use the value that occurs the most frequently i.e. by mode to fill in the blanks.

There are following three ways as given below:

1. Deleting Rows with missing values
2. Impute missing values
3. Prediction of missing values[23]

In this work, the missing values are being imputed for better performance of the machine learning models.

Handling the Outliers

BoxPlot has been used to identify outliers in a dataset. The outliers may be dealt with by either limiting the data or transforming the data[24]:

- *Capping the data*: There are three ways to set data cap limitations this time around.
- *Z-Score approach*: Outliers are any values that go outside of the normal range by a factor of three or more.

CHAID

Decision trees have been around for a long time, but CHAID is the oldest one. Gordon V. Kass first mentioned the issue back in '80. Chi-square is used to determine the relevance of a feature in this case. The greater the statistical significance, the greater the value. CHAID uses decision trees to solve categorization issues in the same way as the others. This implies that a category target variable is expected in the data sets[25]. While ID3, C4.5, and CART all employ information gain, CHAID uses chi-square testing to determine which characteristic is the most prominent. Karl Pearson introduced chi-square testing because of their high accuracy, stability, and simplicity of interpretation. Tree-based learning algorithms are among the finest and most extensively used supervised learning methods[26]. The chi-square statistic is used in the CHAID decision tree technique to determine the independent variable with the biggest chi-square value for the dependent variable. The manifestations of the dependent variable that has the most impact on the dependent variable become the new dependent variable[27]. The individual node in figure 1 consists of category (0 or 1), % (accuracy level) and n (no of patients)[28](Figure 1).

In the study presented here, the compound 'direct bilirubin' is adjudged as most significant factor by the CHAID test, as shown in figure 1, splitting the tree into three sub trees depending upon the quantity of bilirubin found. It also shows the adjusted p-values and chi-square value calculated at each level marking the significant difference between the corresponding sub categories. Further, on the next level, the tree splits on the basis of next important factor, 'alkaline phosphate', for category represented by node 1, i.e., people with value less than equal to 0.9 for compound direct bilirubin and, 'age' for node 2 which represents the direct bilirubin range of 0.9 to 4.1 in a body. The maximum height of the tree allowed is up to the level 5 as the further splits do not significantly affect the result of the model. The significance value kept for splitting the records is .05, using Pearson likelihood ratio for chi-square and Bonferroni method for auto adjusting the significance value and actual p-values. The threshold value for stopping the growth of tree is minimum 2% records in parent branch and 1% record in child branch[29].

Classification & Regression Trees

The C&RT algorithm uses several different ways to divide or segment data into smaller subsets depending on the different values and combinations of predictors that are available. Splits are selected, and the procedure is repeated until the finest possible collection is discovered[30]. Binary splits lead to terminal nodes that may be characterised by a collection of rules, resulting in a decision tree. In order to benefit from the tree's visual appeal and easy-to-understand layout, you don't have to be an expert data scientist[31].

In this study, C&RT produced the binary tree classification for continuous variables and exhibited different sequence by adjudging compound total bilirubin as the least impure predictor[32]. The split-up value for total bilirubin was calculated to be 1.650 by the C&RT model at the first level, using Gini impurity index method, with a gain of 0.047. Subsequently, the other important predictors are 'aspartate_aminotransferase', 'direct bilirubin' and 'age' of the subjects, according to this model. The compound 'total bilirubin' was used at multiple levels for the split which means that this compound produced maximum gain in gini index at multiple levels. The minimum value for recording change in impurity and making a split is set to '0.0001', after a series of run to maximize the efficiency of the model[33]. The maximum number of levels allowed here is '5' in the C&RT tree with the same criteria for stopping the growth as used in CHAID model[34] (Figure 3). An innovative, highly desirable technique that incorporates automation, ease of use, performance and accuracy is what sets CART apart in the predictive analytics sector[35].

Ensemble with Decision trees

The ensemble learning is used in techniques, such as boosting, for repeatedly training the model on different random samples of the dataset. Ensemble methods used in various advanced learning techniques, such as classification & regression trees, XGBoost decision tees, etc., also uses this approach to minimize the error and improve the accuracy[36]. The accuracy level of decision tree may be enhanced by using the ensemble approach (Figure 4).

If the patient have data, the patient may use a tree-like graph to model the choices, which can be either continuous or categorical. As the patient answer each question, the patient will get a forecast about the data that's in front of the patient[37].

XGBoost

XGBoost is a collection of open-source functions and steps that uses supervised machine learning to estimate or predict a result. A result may be predicted using several decision trees in the XGBoost library. Batch learning is used to train the ML system, and then a model-based technique is used to generalise the results. Models for the link between predictor and outcome variables are constructed using all available data. These models are then generalised to the test data.

This stands for eXtreme Gradient Boosting, or simply XGBoost. In the context of computing, the term "extreme" implies its desire to push the boundaries of processing power. Although the concept of "gradient boosting," used to improve the performance of weak prediction models, is used in machine learning applications such as regression and classification[38].

Boosting

Because it can only predict the outcome variable slightly better than chance, a single decision tree is regarded as a weak or basic learner. Strong learners, on the other hand, are any algorithms that can be fine-tuned to attain maximum performance in supervised learning. In order to build a powerful learner, XGBoost employs decision trees as its foundation learners. If you use many models (trees), the final prediction is called an ensemble learning approach since it incorporates the results of multiple models (trees)[39].

It's called "boosting" when a group of weak learners is combined to make a strong learner. Each weak prediction is weighted according to how well the weak learner performed, and XGBoost will repeatedly create a collection of poor models on subsets of the data. The weighted total of all base learners is used to make a prediction[40].

Building models with XGBoost

While all the other features are utilised as predictors of the target variable yi in the training data, the target variable is provided. Decision trees are used to predict the values of yi based on xi using a set of trees. It would be difficult to forecast the result variable with just one decision tree. Analysts may be able to generate more accurate forecasts of yi if the decision trees are used collectively[41].

The learning process of an algorithm is governed by hyperparameters, which are specific values or weights. XGBoost, as previously mentioned, offers a wide range of hyperparameters. XGBoost's hyperparameters may be fine-tuned to reach the highest level of accuracy possible. Auto-tuning of numerous learnable parameters allows for the XGBoost to recognise patterns and regularities in the datasets it analyses. The learnable parameters in tree-based models like XGBoost are the decision variables at each node. A sophisticated algorithm like XGBoost has a lot of design choices and hence a lot of hyperparameters[42].

The main challenge face in this stage is the optimized selection of parameters among multiple hyperparameters. It may be managed by efficient hyperparameter tuning. In this paper, the Bayesian optimization is being applied in following 4 steps as given below:

Step 1. Initialize domain space for range of values

Initially, the domain space is being finalized the input values over which is being searched. The input variables are max_depth, gamma, reg_alpha, reg_lambda, colsample_bytree, min_child_weight and n_estimators.

Step 2. Define objective function

In next step, the objective function is defined, which can be any function that returns a real number that has to decrease in order to achieve the desired goal. XGBoost's validation error with respect to hyperparameters should be minimised in this situation. In order to optimise accuracy, the other key value must be considered. It should return a value that's the opposite of that metric's value.

Step 3. Apply Optimization algorithm

It is the method used to construct the surrogate objective function and choose the next values to evaluate. In this stage, this paper employ the concept of Bayesian Optimization in this tuning phase. The Bayesian Optimization method is based on the Bayes Theorem and provides an efficient and effective method for solving a global optimization issue. The real objective function is then used to evaluate the candidate samples.

Step 4. Results

Results are score or value pairs that the algorithm uses to build the XGBoost model.

For XGBoost model, the objective function optimization is performed using the logistic model, as the target is a categorical variable. The tree building model used is 'auto' with ten iterations for boosting at each level. The tree depth, with which maximum efficiency is obtained, is found to be up to level 6, wherein, minimum child weight is set as default '1' and maximum delta step is set to no constraint with value '0' (Figure 5). Similarly, the parameter for column and level sampling are also set to value '1' (Figure 6).

The value for least square error and least absolute deviation represented by parameter lambda and alpha, respectively, is set to value '1' and '0' which is also used for regularizing the weights, make the model more stable and control the overfitting[36] (Figure 7).The model is also executed with hyperparameter tuning setup using Bayesian optimization model and showed even better performance than the non-hyperparameter tuning setup.

RESULTS

Results from Individual Decision Trees

The output of the different prediction models, presented in this work, is shown in this section in the categorized tabular format. The Gini Coefficient is a machine learning metric used to assess the efficiency of binary classifier models. In the range of 0 to 1, the Gini Coefficient can be used. The higher the Gini coefficient, the better the model. There are two ways to measure precision: the number of properly categorized positive samples (True Positives) and the overall number of positively classified samples (either correctly or incorrectly). Using precision, we can see how reliable the machine learning model is when it comes to determining whether or not the model is positive

The recall is computed by dividing the total number of Positive samples by the number of Positive samples that were correctly categorized as Positive. The recall assesses the model's capability to identify positive samples in a data set. The more positive samples are found, the higher the recall. Figure 8-10 depicts the accuracy of an individual decision tree and tabulates the values of various performance matrices (Figure 8).

There are a total of 583 patients in the Indian Liver Patient Dataset. A total of 500 patient records were utilised in the training process, and an additional 83 records were used in the testing process. The model produced the Gini coefficient value of 0.49 exhibiting good discriminatory behavior and showed 71.36% accuracy in correctly classifying that whether a patient has liver disease or not based on the input data (Figure 9).

Out of the 583 observations, the predictions can be categorized as 389 to be true positive, 38 as true negative and 156 as false predictions in this model. The Gini coefficient for the model is calculated as 0.44 which shows that the ability of the model to categorize is not good enough (Figure 10).

The multiple machine learning methods for the prediction of liver disease are assessed in this study (Table 2).

Table 2 depicts accuracy visually along with AUC & Gini value. The figure 11 highlights the result analysis in graphically manner as given below (Figure 11).

DISCUSSION

Artificial Intelligence plays a very important role in predicting liver disease. It consists of different machine learning algorithms which may be applied in liver disease prediction. In this paper, multiple machine learning algorithms have been implemented on the above said dataset of Indian liver patients. This work consists of the execution of CART and CHAID algorithms to predict the liver patients. If bilrubin, protein, alkaline phosphatase, and albumin are present in the human body, and tests like SGOT and SGPT indicate that a person needs to be diagnosed, the results are stated through the various machine learning algorithms. The proposed algorithm is observed as efficient in handling the missing values and overfitting problem. It is tested as capable of handling high variance problems normally faced in execution of the machine learning algorithms.

CONCLUSION

The decision tree algorithm i.e. CART & CHAID are found not such more accurate. This factor motivate the authors to design the proposed machine learning model (Hybrid XGBoost model) which gain the accuracy level of 93.65%. This proposed model faced the over-fitting issue during its execution phase. In current execution, the textual dataset of 563 patients has been used. In future, the image dataset of lung disease should be taken for avoiding other tests like SGPOT & SGPT. Such approaches are very useful for minimizing the workload of the doctors.

ARTICLE HIGHLIGHTS

Research background

Liver disease is a leading cause of mortality in the United States and is regarded as a life-threatening condition not only in this nation but all across the world. It is possible for people to develop liver disease at a young age.

Research motivation

Predicting liver illness with more precision, accuracy, and reliability can be accomplished through the use of a modified XGBoost model with hyperparameter tuning in comparison to CHAID & CART models.

Research objectives

This paper has been written with aiming of fulfilling various objectives. First objective of this paper is identifying the symptoms of the liver disease and its impact on the patient’s body. Then this paper studies various machine learning approaches for predicting liver disease and evaluate the performance of decision-tree algorithms in prediction of liver disease. Next objective of this paper is concerned to propose a modified XGBoost model with hyperparameter tuning mechanism. Finally it has validated the performance of the proposed model with the existing model.

Research methods

Hybrid XGBoost model with Hyperparameter Tuning

Research results

The CHAID & CART model have achieved the accuracy level 71.36% & 73.24% respectively. With the help of the proposed model, the accuracy level has been achieved up to 93.65%.

Research conclusions

The CHAID & CART model have achieved the accuracy level 71.36% & 73.24% respectively. With the help of the proposed model, the accuracy level has been achieved up to 93.65%.

Research perspectives

As a result, it helps patients to predict liver-related disease acutely by identifying the disease's causes and suggesting better treatment options.

78509_Auto_Edited.docx

ORIGINALITY REPORT

2%

SIMILARITY INDEX

PRIMARY SOURCES

1	www.kaggle.com Internet	30 words — 1%
2	dspace.bracu.ac.bd:8080 Internet	27 words — < 1%
3	resources.rhoworld.com Internet	13 words — < 1%
4	consumer.healthday.com Internet	12 words — < 1%
5	diabeticnewsletter.com Internet	12 words — < 1%

EXCLUDE QUOTES	ON	EXCLUDE SOURCES	< 12 WORDS
EXCLUDE BIBLIOGRAPHY	ON	EXCLUDE MATCHES	< 12 WORDS